Exploring the Heterogeneous Impact of Immigrants on US labor Market Outcomes

Ruxin Chen*

June 6, 2018

Abstract

The effect of immigrants on the employment opportunities for native workers has received a great deal of research attention during the last decades. Previous empirical studies regarding this question do not reach a consensus and there is a lack of a widely-acknowledged theoretical model to explain the interplay of immigrants and native workers. Using US census data, the paper applies a regression-tree based method proposed by Athey and Imbens [2016] to identify the sub-population and the heterogeneous effect of immigrants on labor market outcomes. The findings suggest that the proportion of immigrants have insignificant effect on all subgroups of US labor force, however, it affects the individuals with graduate degree to a larger magnitude than individuals without graduate degree in terms of their wage income.

*University of Chicago, Social Science Division, Master of Computational Social Science ruxin@uchicago.edu.

1 Introduction

The effect of immigrants on the employment opportunities for native workers has received a great deal of research attention during the last decades. In recent years, as the intensifying of the regional terrorist threats in the Middle East, whether to accept refugees has been at the heart of a growing political debate for many western countries. Moreover, with the slowing pace of economic growth in the developed world, policy makers seem to be concerned about the adverse effect, if any, on employment that immigrants might bring to the native labor market. My project intends to investigate this question using a computational method to add value to the literature. Specifically, it will deploy a regression tree method to identify the heterogeneous effect of immigration on native labor market outcomes on different subgroups of the population. This study is expected to provide some policy implications under the current political landscape of exclusionism/protectionism.

2 Literature Review

The classical model suggests that an inflow of immigrants would always lower the wage of competing factors. Though the model is well understood, the economic mechanism of the interplay of immigrants and native workers might go beyond the over-simplified demand-supply model. According to Friedberg and Hunt [1995], the theoretical prediction of this problem will depend on the whether the economy of the host country is closed or open and also the composition of the native workers and immigrants in terms of their skills set: "In a closed economy, immigrants will lower

the price of factors with which they are perfect substitutes, have an ambiguous effect on the price of factors with which they are imperfect substitutes and raise the price of factors with which they are complements." Hence, skilled immigrants will lower the skilled wage, increase the employment in the skilled occupations, and the result is deducted likewise for the unskilled immigrants. However, in an open economy, the Heckscher-Ohlin model indicates that countries specialize in the production of goods according to their initial endowments of factors. Labors will migrate from abroad as long as the domestic wage is above the world level. Therefore, with certain restrictions on immigration, the wage level will rely on the size: if the number of immigrants allowed is much larger than the equilibrium at which the immigration fully eliminates the wage differentials, the country will move to more labor-intensive goods which lowers the wage; if there are few immigrants, the wage level remains the same as the immigration effect can be digested through exporting the excessive goods. While the case is different when we try to explain by the efficiency wage model of Shapiro and Stiglitz [1984]. It is hard to determine the optimal theory that should be applied in this case by logical reasoning, hence, a careful empirical study is valuable to improve the theoretical framework.

However, the previous empirical results did not come to a consensus regarding this question. The vast majority of the existing studies conclude that the impact of immigration on the labor market outcomes is small. LaLonde and Topel [1991] examined that despite the fact that the immigrants brought fewer marketable skills and were less-educated, they assimilated rapidly in the US market during the 1970s and 1980s and had a negligible impact on the wages and employment prospect on the native

workers. Council et al. [1997] found similar results in the 1997 National Academy of Sciences report. However, Card [1990] argued that the presence of spatial correlation would cause bias in estimation. According to Borjas [2003], papers that use the differences between labor markets with/without clustering of immigrants to identify the impact of immigration ignore the fact the mobility of native labors will attenuate the effect. Moreover, the endogeneity problem might arise because immigrants tend to concentrate in big cities where the wage level is higher per se. He then proposed that, instead of using geographical variation, an effect can be estimated by dividing the population of interest by different skills group based on education attainment and experience, and for each skills group, compare the differences between those with and without an immigrant supply shock. He obtained an adverse effect of immigrants on the native workers: a 10 percent increases in supply reduces wage of native workers by 3 to 4 percent. Similarly, Angrist and Kugler [2003] also examined a negative effect of immigration on the native workforce, but to a different magnitude, where a 10 percent increase in supply would reduce native employment rate by 0.2-0.7 of a percent point.

Based on the argument of "spatial correlation", Dustmann et al. [2005] propose an identification strategy by applying pre-existing immigration concentration as an instrument for the current immigration shock. They used the data from British, where unlike the US and some continental European countries where the immigrants are often less-skilled, the composition of the British immigrants is quite similar to the composition of the native workforce. Interestingly, they found no overall adverse effects of immigration on native outcomes. The negative effect for the group with

intermediate education levels is offset by the positive effects on employment for the better-qualified groups. A more recent paper by Foged and Peri [2016] considers the refugee dispersal policy as a quasi-experiment to obtain identification. Different from all previous study, the author reported a positive effect on the less-educated native workforce in terms of wage, employment, and occupational mobility.

Both the theories and empirical results imply that the effect of immigration on native worker might be heterogeneous across different groups. Understanding the heterogeneity might help us to figure out the reason for the inconsistency of the findings in the literature. Existing papers only focused on one particular type of heterogeneity: high-skilled and low skilled [Borjas, 2005, Athey and Imbens, 2016]. However, the definition of skilled and unskilled seem to be ambiguous. Borjas [2003] considers both the education attainment and work experience to sort workers into particular skilled groups. However, as acknowledged by the author, "the classification of workers into experience groups is bound to be imprecise because the Census does not provide any measure of labor market experience or of the age at which a worker first enters the labor market". The measure he applied to approximate the experience, "age - education -6", is biased for female workers and cannot differentiate between experience obtained in the source country and the United State. More importantly, it is dubious that we can consider workforce with similar skills level as substitutes in the labor market. An apparent example is that employers will be likely to treat differently for young employees and old employees (though age is correlated with experience, age definitely deliver information other than experience to employers). Other factors like age might also be essential to heterogeneity problem but are failed

to captured in theories. Therefore, a machine learning technique is applied to identify the potential subgroups of the population. This is the major innovation of my study. The research is deductive, and the finding is expected to provide some insights to address the confusion in theories.

The machine learning method I am going to use is based on the method proposed by Athey and Imbens [2016], which uses a regression tree to find the partition of the population according to covariates. Most machine learning technique cannot be used directly for constructing confidence interval since the models are "adaptive" and the conventional asymptotic properties cannot be achieved. They propose a two-steps procedure to overcome this problem, called "honest" approach, which splits the training sample into two parts for constructing the tree and estimating treatment effect within each leave of the tree. The independence of the two steps guarantees the validity of the estimation. The detailed description of the methodology will be presented in the following sections.

3 Data

The IPUMS preserves U.S. census microdata from 1790 to 2010 and American Community Surveys (ACS) from 2000 to the present. The database contains large samples of individual demographic characteristics, geographical location, labour market outcomes that are sufficient for our study. We initially requested a data set consists of more than 3,800,000 observations from 2001 - 2016, which occupies about 4 GB of space. The data are formed by a 1-in-100 random sample of the population. Variables

include state, county, metropolitan status, age, sex, marital status, race, ethnicity, year of immigration, language spoken, education attainment, labor force status, employment status, occupation, class of worker, usual hours worked per week, place of work in terms of state and wage and salary income. We create dummy variables to denote some of the above-mentioned categorical data. Since we are interested in the changes of labor market outcomes, we drop the observations that are not in the labor force. In the first part of our analysis, we study the effect to US labor market outcomes caused by the fraction of immigrants by county, hence, we cluster the data by county and year, so that the categorical data are denoted as the proportion of population in each group. For example, a dummy variable "Female" is created in the raw data indicating the sex of the individual, in the clustered data, we calculate the mean of this variable, which is simply the proportion of female in the associated county and year. The clustered data have 4450 observations in total. In order to explore the heterogeneous effect of immigrants, since a great deal of important information is masked by clustering, we switch back to individual level data. Given the large size of our data, we randomly draw 5% from the 2006 data. The number of observations is 73888 and it is suffice for our analysis.

Admittedly, our data set is not comprehensive since it includes only around 600 counties for each year. However, the ACS data is the only data we can find that include both labor market outcomes as well as individual demographics which are essential for our study. Figure 4 shows the proportion of immigrants on map from 2006 data. The blank areas are those counties with missing data. Our data are distributed in various regions in US, though it seems to be concentrated on the eastern

area and sporadic in the middle-north. This pattern is close to the natural population density in US, and therefore cause less concern to the validity of our results.

Next, table 1 reports the summary statistics for some selected variables of interest. The average proportion of immigrants is 12%. The distribution is right-skewed, which suggests that we have more counties with relative small percentage of immigrants, while very few counties have immigrants more than half of its entire population. During 2006-2015, the unemployment rate is about 7% with a standard deviation of 0.03%, the shape of the histogram is close to a normal distribution except that it appears to be a bit right-skewed. In our sample, the majority of the population (54%) entered in colleges, on average, 20% of the population obtained a college degree and 12% have advanced degrees.

Figure 1: Proportion of Immigrants in US in 2006

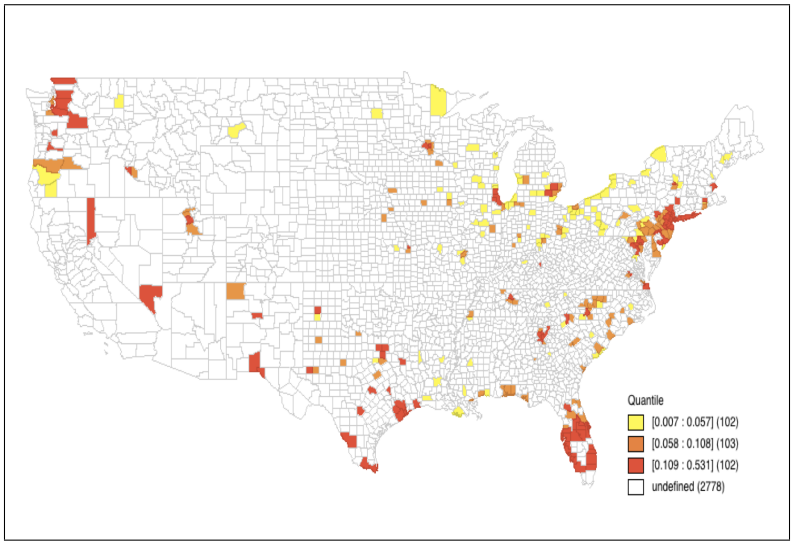


Figure 2: Histogram of Proportion of Immigrants for US Counties during 2006-2016

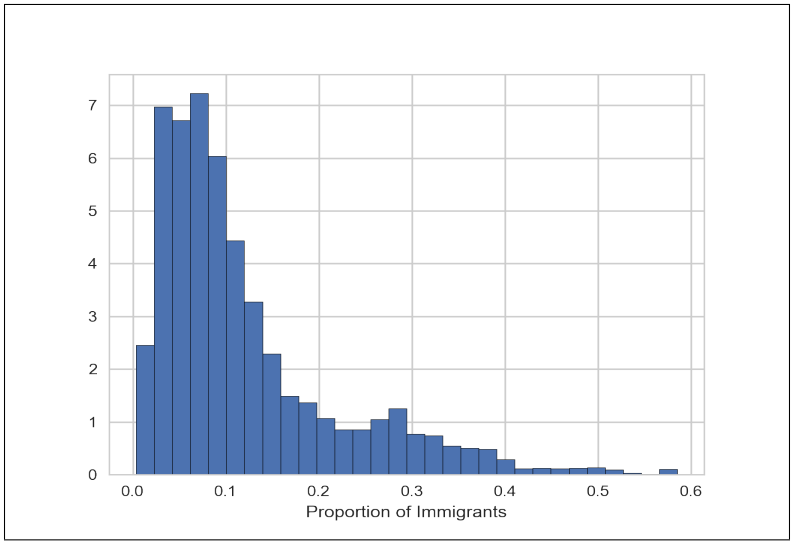


Figure 3: Histogram of Average Wage Income for US Counties during 2006-2016

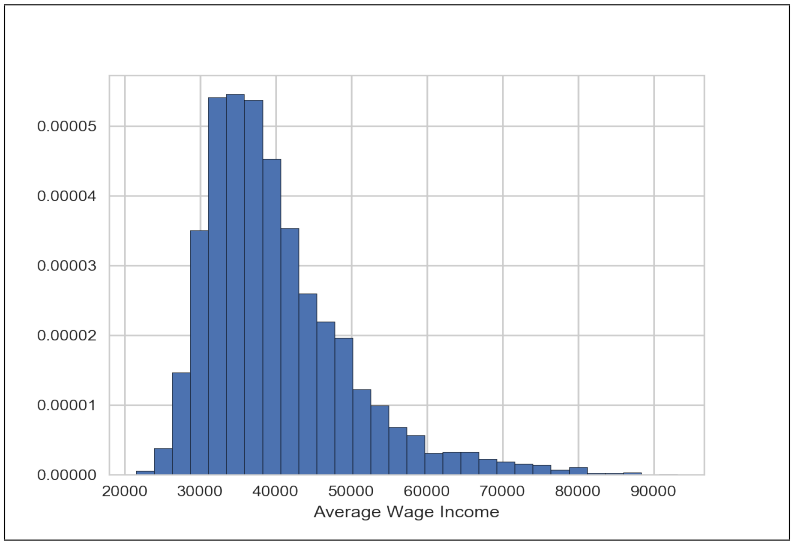


Figure 4: Histogram of Unemployment Rate for US Counties during 2006-2016

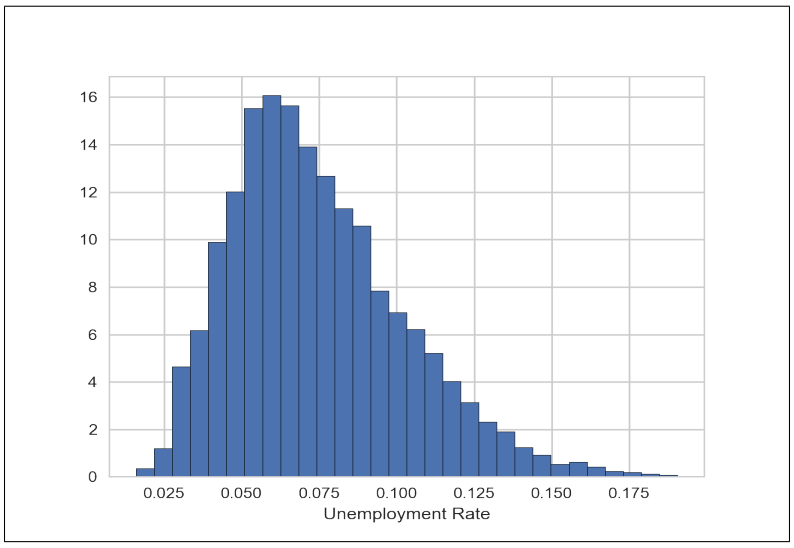


Table 1: Summary Statistics for Selected Variables for Clustered Data

	Mean	Standard Deviation
Proportion of Immigrants	0.12	0.10
Unemployment Rate	0.07	0.03
Usual hours worked per week	37.56	1.57
Wage and salary income	40129.10	9712.91
Average Age	42.51	1.68
Proportion of Female	0.48	0.02
Proportion of White	0.76	0.29
Proportion of Black	0.09	0.11
Proportion of Hispanic	0.10	0.13
Proportion of Elementary	0.00	0.00
Proportion of Some College	0.54	0.14
Proportion of College Degree	0.20	0.06
Proportion of Advanced Degree	0.12	0.05
Proportion of No Education	0.01	0.01
Proportion of High School Degree	0.14	0.13
Observations	4450	

4 Methodology and Model

4.1 Preliminary Results: OLS and Fixed Effect Estimation

We first specify the regression model to describe the relationship between the fraction of immigrants on US labor market outcomes as follows:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \delta C_{it} + \gamma_t + \eta_i + u_{it} \tag{1}$$

where y_{it} is the outcome of interests including the unemployment rate and the average income wage for county i in year t. x_{it} denotes the proportion of immigrants. C_{it} is a set of observable characteristics which consist of the average age for the county, proportion of female, proportion of white/black population, proportion of Hispanics, and the education attainment level. The education attainment is recorded as categorical data, hence we classify individuals into: no education, elementary, high school, some college, college degree, advanced degree(graduate degrees), and include the fraction of each category into our regression. γ_t and η_i denote the year and state fixed effect respectively. Then we estimate the model with simple OLS method and a fixed effect method.

Table 2 presents the regression results for unemployment rate using OLS estimation, year-fixed effect model and year-state fixed effect model. The coefficient for proportion of immigrants are significant for naive OLS estimation and model with year fixed effect. However, after taking account into the state fixed effect, the effect of immigrants becomes insignificant. Education attainment, race and gender compo-

sition seem to be important factors to the analysis of the employment status, which is consistent with theories in labor economics. Notice that except for the proportion of elementary school, all other categories of education level are significant for all the three models. The result makes sense as the proportions of individuals have only finished elementary school are low for all counties in US. Table 3 reports the estimation results for average wage income. Unlike the unemployment rate, the proportion of immigrants have significant positive effect on the average wage income for all labor force across different model specifications. While, the wage income increases as age increases, and female generally earn less than male in the labor market.

In the next section, we will demonstrate how to identify different subgroups that are affected heterogeneously by the immigrants. This innovation is essential in interpreting the coefficients. As mentioned in the literature review section, a "zero" coefficient might be caused by either homogeneous insignificant effect for all of the subgroups or adverse effects for different subgroups with a zero sum. The policy implication might different dramatically with the two different scenarios.

Table 2: Model Estimation Results for Unemployment Rate

	(1) OLS	(2) Year Fixed Effect	(3) Double Fixed Effect
Proportion of Immigrants	0.0897***	0.0739***	0.0052
	(0.01)	(0.01)	(0.01)
Average Age	0.0002	0.0001	-0.0007***
	(0.00)	(0.00)	(0.00)
Proportion of Female	0.2545***	0.2278***	0.2052***
	(0.02)	(0.02)	(0.01)
Proportion of White	0.0173***	-0.0011	-0.0102
	(0.00)	(0.00)	(0.01)
Proportion of Black	0.0607***	0.0407***	0.0536***
	(0.00)	(0.01)	(0.01)
Proportion of Hisp	-0.0092*	-0.0078*	0.0263***
	(0.00)	(0.00)	(0.00)
Proportion of Elementary	0.2212	0.1379	-0.2613*
	(0.17)	(0.14)	(0.12)
Some College	-0.2192*	-0.1928*	-0.2589***
	(0.10)	(0.08)	(0.07)
College Degree	-0.4328***	-0.3758***	-0.3795***
	(0.10)	(0.08)	(0.07)
Advanced Degree	-0.2694**	-0.2269**	-0.2958***
	(0.10)	(0.08)	(0.07)
Proportion of High School	-0.2895**	-0.1947*	-0.2353***
	(0.10)	(0.08)	(0.07)
Constant	0.1882	0.1640	0.2655***
	(0.10)	(0.08)	(0.07)
R-squared	0.4170	0.5952	0.7339
N. of cases	4450	4450	4450

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Model Estimation Results for Average Wage Income

	(1) OLS	(2) Year Fixed Effect	(3) Double Fixed Effect
Proportion of Immigrants	15132.2934***	19611.4329***	17277.9796***
	(1158.14)	(1262.44)	(1445.59)
Average Age	1359.9532***	1319.4162***	1324.0119***
	(41.14)	(40.83)	(40.57)
Proportion of Female	-105279.1304***	-105074.1981***	-93461.1732***
	(3724.24)	(3644.05)	(3422.84)
Proportion of White	-2485.0698***	4685.9967***	-632.2032
	(229.99)	(1056.58)	(1568.11)
Proportion of Black	2224.9170**	8533.9001***	-490.1091
	(697.92)	(1197.09)	(1695.79)
Proportion of Hisp	-4701.5463***	-6884.2696***	-8424.4103***
	(913.66)	(919.63)	(1086.55)
Proportion of Elementary	-29458.3118	2978.5478	17632.7882
	(33758.85)	(33655.73)	(29689.06)
Some College	13157.5449	48179.7380*	44668.2618**
	(19126.67)	(19430.32)	(17289.19)
College Degree	84810.3011***	121651.8367***	123162.5249***
	(19009.72)	(19420.38)	(17217.66)
Advanced Degree	92398.9611***	128261.8450***	108429.5133***
	(18997.49)	(19325.58)	(17280.39)
Proportion of High School	11524.2841	55898.7708**	52259.4448**
	(18945.63)	(19663.31)	(17480.97)
Constant	-2830.9091	-47928.3462*	-44704.0374*
	(19118.59)	(19596.96)	(17479.36)
R-squared	0.8105	0.8201	0.8683
N. of cases	4450	4450	4450

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.2 Decision Tree Based Method

4.2.1 Conventional Average Regression Tree (CART) and "Honest" Approach

According to the results for the previous section, we will determine the division of the subgroups based on the observable characteristics, C_{it}. Regression tree is an unsupervised machine learning technique that differs from most other methods in that they produce a partition of the population according to covariates[Athey and Imbens, 2016]. The CART algorithm recursively partitions the observations of the training sample. For each leaf, the algorithm evaluates all candidates splits of that leaf using a "splitting" criterion that is referred as the "in-sample" goodness-of-fit criterion

$$\min_{j,s} MSE = \min_{j,s} \left(\min_{\mu_1} \sum_{x_i \in R_1(j,s)} (y_i - \mu_1)^2 + \min_{\mu_2} \sum_{x_i \in R_2(j,s)} (y_i - \mu_2)^2 \right) \tag{2}$$

where, consider a splitting variable j and a splitting point s such that

$$R_1(j,s) = \{X | X_j \leq s\} \quad \text{and} \quad R_2(j,s) = \{X | X_j > s\} \tag{3}$$

where a minimization solution of of μ_i for $i = 1, 2$ is the conditional sample average for y_i:

$$\hat{\mu}_i = \frac{1}{\#\{i \in n : x_i \in R_i(j,s)\}} \sum_{i \in n : x_i \in R_i(j,s)} y_i \tag{4}$$

However, the CART approach is "adaptive" in the sense that the same training sample is used to construct and estimate the regression tree. As a consequence, the estimation of parameters does not follow the asymptotic theory as the standard regression estimation method: the biases only disappear slowly as sample size grows. In order to obtain an inference, Athey and Imbens [2016] proposed the idea of causal tree and modified the "adaptive" method to be an "honest" one by imposing a separation between constructing the partition and estimating effects within leaves of the partition, using separate samples for the two tasks, denoted as S^{tr} and S^{est} respectively. The "honest approach" avoids the spurious correlation between the model selection to the estimation process at the cost of sacrificing the size of data we use for estimation, because we need to take out some data to construct the regression tree first. However, given the large amount of data we collect, this partition should not affect the validity of the results as long as the model is appropriate. Athey and Imbens [2016] discovered that, if we split our data equally for training sample and estimation sample, the criterion function of the "honest" approach equals

$$EMSE(S^{tr}, s, j) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\mu}^2(X_i; S^{tr}, j, s) - \frac{1}{N^{tr}} \sum_{x_i \in R_i(j,s)} S^2_{S^{tr}}(R_i(j, s)) \quad (5)$$

where the first term is simply the criterion function for CART, and the second term is the summation of within-leaf sample variances estimated by the training sample. In the next section, we will present the results of our regression model after modifying the criterion function algorithm as illustrated above.

4.2.2 Estimation Results

In this section, we use a random sample of individual data in 2006 to construct the tree and do the estimation. The model specification for examining the wage income effect is

$$y_i = \beta_0 + \beta_1 x_i + \delta C_i' + \eta_i + u_i \tag{6}$$

where y_i is the wage income, x_i is the proportion of immigrants for the county where individual i resides, C_i' denotes the a set of dummy variables based on the splits found by regression tree on the observable characteristics, η_i is the fixed effect for the state where individual i resides and u_i is the error. The model for examining the unemployment rate effect is

$$P(y_i = 1|x_i, C_i') = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i + \delta C_i' + \eta_i + u_i)}} \tag{7}$$

where $P(y_i = 1|x_i, C_i')$ denotes the probability of being unemployed for individual i, and the definition of the remaining variables follows those for equation 6.

Figure 6 and 5 show the estimation results using the causal tree method. We set the max depth of the tree to be two.

We found the population will be split for individuals with and without a graduate degree. In general, the wage income will be increased as the proportion of immigrants increases for all subgroups of the population. The magnitude of the effect is found to be larger for individuals with advanced degree/graduate degree. Among people with graduate degrees, the male are likely to earn higher in a county with more immigrants.

However, the individuals with college degree with be influenced to a smaller extent to those with no college degrees. The probability of being unemployed is estimated by a logit regression. According to figure 6, though the splits are detected for population with different education attainment and gender and the signs of the estimated effects are different, none of the effect is significant.

Figure 5: Decision Tree Estimation Result for Wage Income

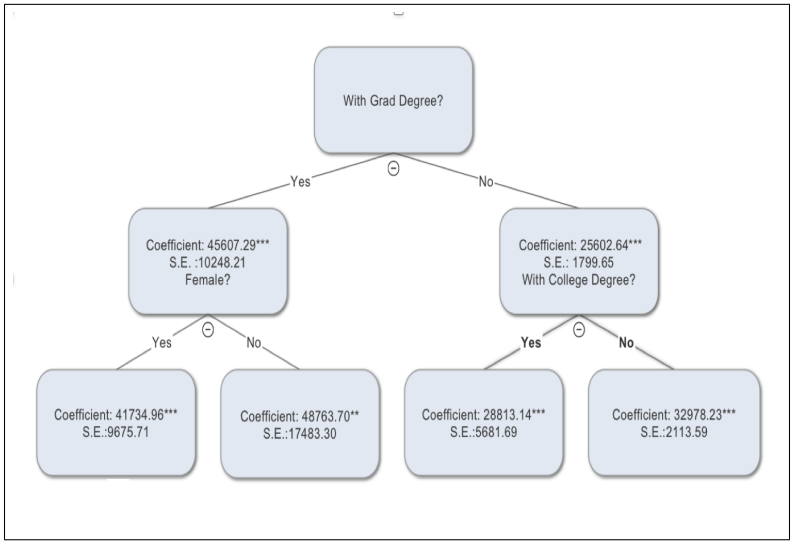
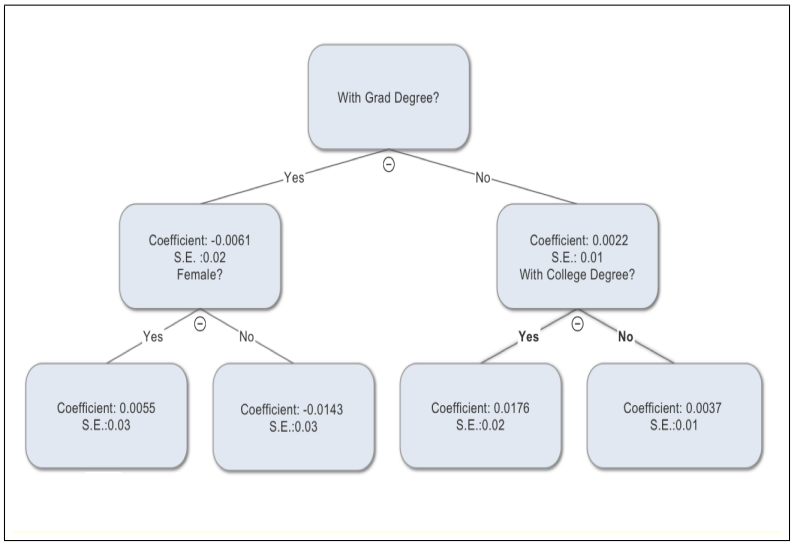


Figure 6: Decision Tree Estimation Result for Unemployment Rate



5 Limitations and Future Research

Though efforts have been put to figure out an exogenous variation to identify the causal effect of immigrants on the US labor market, the initially proposed quasi-natural experiment of the Secure Fence of Act in 2006 is examined to be actual endogenous using the data, and further investigation is limited due to the lack of precise geographical location for individual before 2006. Hence, we are cautious to claim the findings in this study to be causal. Another concern for this paper is that, we dropped the non labor-force from our data for convenience. However, the wage income should be a latent variable that is only observable if people choose the participate into the labor market. The determination is based on the relative magnitude of the anticipated wage if participate and the reversed income (opportunity cost for participating), which might also be affected by the proportion of immigrants. Future study can be conducted by 1). proposing an identification strategy 2). specifying the

wage income as a latent variable and estimate with a discrete choice model.

6 Conclusion

The paper investigates the relationship between the proportion of immigrants and the labor market outcomes such as unemployment rate and wage income. Panel data are used to estimate the effect. The average wage income is affected positively as the proportion of immigrants within such county is increased, under both the OLS and fixed effect specifications. The effect on unemployment rate is significantly positive if only year fixed effect is considered and it is disappeared if we add a state-fixed effect in to the model. In the second part of the paper, we identify the sub-population that are influenced heterogeneously by immigrants in terms of their labor market outcomes using a regression tree based method. This method is called "honest approach" that address the inconsistency of estimation for adaptive models. The findings suggest that the proportion of immigrants have larger effect on individuals with more advanced degree, and among people with graduate degrees, the male are likely to earn higher in a county with more immigrants. Unlike the wage income, the probability of being unemployed seems to be not affected by the proportion of immigrants.

Bibliography

Angrist, Joshua D and Adriana D Kugler, "Protective or counter-productive? Labour market institutions and the effect of immigration on EU natives," *The Economic Journal*, 2003, *113* (488).

Athey, Susan and Guido Imbens, "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 2016, *113* (27), 7353–7360.

Borjas, George J, "The labor demand curve is downward sloping: Reexamining the impact of immigration on the labor market," *The quarterly journal of economics*, 2003, *118* (4), 1335–1374.

_ , "The labor-market impact of high-skill immigration," *American Economic Review*, 2005, *95* (2), 56–60.

Card, David, "The impact of the Mariel boatlift on the Miami labor market," *ILR Review*, 1990, *43* (2), 245–257.

Council, National Research, Committee on Population et al., *The new Americans: Economic, demographic, and fiscal effects of immigration*, National Academies Press, 1997.

Dustmann, Christian, Francesca Fabbri, and Ian Preston, "The impact of immigration on the British labour market," *The Economic Journal*, 2005, *115* (507).

Foged, Mette and Giovanni Peri, "Immigrants' effect on native workers: New analysis on longitudinal data," *American Economic Journal: Applied Economics*, 2016, *8* (2), 1–34.

Friedberg, Rachel M and Jennifer Hunt, "The impact of immigrants on host country wages, employment and growth," *Journal of Economic perspectives*, 1995, *9* (2), 23–44.

LaLonde, Robert J and Robert H Topel, "Immigrants in the American labor market: Quality, assimilation, and distributional effects," *The American economic review*, 1991, *81* (2), 297–302.

Shapiro, Carl and Joseph E Stiglitz, "Equilibrium unemployment as a worker discipline device," *The American Economic Review*, 1984, *74* (3), 433–444.